Further results on latent discourse models and word embeddings

Sammy Khalife, Douglas S Gonçalves, Youssef Allouah, Leo Liberti

Further results on latent discourse models and word embeddings

Sammy Khalife[1], Douglas S. Gonçalves[2], Youssef Allouah, and Leo Liberti[1]

[1] LIX, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France
[2] MTM/CFM - Universidade Federal de Santa Catarina, 88040-900, Florianópolis, Brazil
khalife@lix.polytechnique.fr douglas@mtm.ufsc.br
youssef.allouah@polytechnique.edu liberti@lix.polytechnique.fr

Abstract. We discuss some properties of generative models for word embeddings. Namely, [Arora et al., 2016] proposed a latent discourse model implying the concentration of the partition function of the word vectors. This concentration phenomenon lead to an asymptotic linear relation between the pointwise mutual information (PMI) of pairs of words and the scalar product of their vectors. Here, we first show that the concentration phenomenon is rather general, since it holds for random vectors symmetrically distributed around the origin. Second, we empirically evaluate the relation between PMI and scalar products of word vectors satisfying the concentration property. Our findings indicate that the relation PMI and scalar product fail to occur with empirical word embeddings. We deduce that either natural language does not follow the assumptions of the generative model, or the current methods do not allow the reconstruction of the hypothesized word embeddings. Finally, we provide necessary conditions for the positivity of the shifted symmetric PMI matrix in terms of local pairwise probabilities and provide evidence that it fails to be positive semidefinite in practice, which shows that a result by [Levy and Goldberg, 2014] does not apply to symmetric models. This implies that the linear relation between PMI and scalar product cannot hold with arbitrarily small error.

1 Introduction

Context and Motivations

The construction of intermediate representations is essential for language models and their applications. These representations can be cast in two groups. First, vector space models with *static* word embeddings, where a minimal unit of the language, a word, is associated to a fixed and constant representation. This representation encodes the meaning of the word, independently of its context. There exist many examples of static representations, such as word2vec [Mikolov et al., 2013a] or Glove [Pennington et al., 2014] representations. The second family, which we refer to as *contextual* embeddings, maps each word of the vocabulary to a vector which depends on its context. Long short term memory (LSTM) networks [Hochreiter and Schmidhuber, 1997], or neural networks with attention mechanisms (e.g. Bidirectional Transformers [Vaswani et al., 2017,Devlin et al., 2019]) are examples of methods to construct contextual representations.

Despite the fact that contextual embeddings are considered to have superseded the use of standard vector space models for applications, most of their properties, in particular the relation with language semantics, remain obscure. On the other hand, the family of static embeddings in [Arora et al., 2016] have been advertised to possess geometric properties related to language semantics, in particular with respect to analogies. In this work, we discuss the foundations of such statements, in particular concerning the properties of a latent model for natural language generation.

Previous Work

The model we will consider has been presented in [Arora et al., 2016]: a generative model using prior probability distributions to compute closed form expressions for word statistics. It originally aimed at providing a piece of explanation of the linear structure for analogies [Arora et al., 2016,Arora et al., 2018b]. The apparent relation of linear structures of word vectors and semantic analogies has already been studied in [Khalife et al., 2019] going in favor of an incidental phenomenon rather than systematic. For the sake of clarity, we will present in the remaining of this subsection the main assumptions of this generative model.

In the following, $f = O(g)$ (resp. $f = \tilde{O}(g)$) means that f is upper bounded by g (resp. upper bounded ignoring logarithmic factors) in the considered neighborhood. Let d be an strictly positive integer corresponding to the word vectors dimension. The generation of sentences in a given text corpus is made under the following generative assumptions.

- *Assumption 1:* The text generation process is driven by a random walk of a vector c_t, i.e. if w_t is the word at step t, there exists a latent discourse vector c_t such that

$$\mathsf{P}(w_t = w|c_t) \propto \exp(\langle c_t, v_w \rangle) \tag{1}$$

  where $v_w \in \mathbb{R}^d$ is the word vector for word w. Moreover, the random walk $(c_t \,|\, t \geq 1)$ admits a uniform stationary distribution on the unit sphere.
- *Assumption 2:* The ensemble of word vectors consists of independent and identically distributed (i.i.d.) samples generated by $v = s\,\hat{v}$, where $\hat{v}$ is drawn from the spherical Gaussian distribution in $\mathbb{R}^d$ and s is an integrable random scalar such that $|s| \leq \kappa$.
- *Assumption 3:* $(c_t \,|\, t \geq 1)$ jumps are small in average. More precisely, $\exists \epsilon_1 \geq 0$ such that $\forall t \geq 1$:

$$\mathbb{E}_{c_{t+1}}(e^{\kappa\sqrt{d}||c_{t+1}-c_t||_2}) \leq 1 + \epsilon_1 \tag{2}$$

Contributions

The contribution of this work is three-fold. First, we present a theoretical result concerning the concentration of a partition function Z_c for a generative model following Equation (1). Our statement concerns the behavior of the partition function: it concentrates around its mean, for simpler assumptions than those 1 to 3 aforementioned. Informally, this property means that the variations of Z_c with respect to c (and its randomness) are relatively negligible. This property is very similar to the concentration phenomenon demonstrated in [Arora et al., 2016]. Our result suggests that it is not an intrinsic characteristic of word embeddings satisfying the Gaussian prior (Assumption 2), since it holds for other random vectors symmetrically distributed around the origin.

Second, we empirically investigate the relation between the geometry of word vectors and PMI. Our extensive experiments strongly support the claim that theoretical relations derived from the generative models occur at best in some regimes of the co-occurrence terms. Finally, we provide evidence that the underlying implicit matrix factorization problem necessary to construct word embeddings is ill-posed for a symmetric PMI model, since the shifted PMI matrix (as explored in [Levy and Goldberg, 2014]) is not positive semidefinite. To do so, we establish necessary conditions for the positive definiteness of the shifted symmetric PMI matrix in terms of local pairwise probabilities and show these local conditions can be violated in natural language.

2 Result on the Concentration of the Partition Function

In this section, we discuss a theoretical property presented in [Arora et al., 2016], called the concentration of the partition function. Based on (1), given a discourse vector c, the corresponding partition function value Z_c is defined as:

$$Z_c = \sum_v \exp(\langle v, c \rangle) \tag{3}$$

where v are the word vectors. We remind our reader that the considered generative model treats corpus generation as a dynamic process, where the t-th word is produced at step t. The process is driven by a random walk of a discourse vector c. Its coordinates represent the current topic. In this section, we are interested in an asymptotic property of the partition function Z_c. By analogy with statistical physics, this partition function is the sum of probabilities of the particles state given macroscopic parameters, such as temperature, over all the particles. More precisely, in our context, the particles considered are words and the states are the appearances of a word given a latent discourse vector (which is the equivalent of the physical temperature). This latent discourse vector represents a context of fixed length. The aim of this section is to study the variations of Z_c with respect to the random variable c. This study is motivated by the use of partition concentration as a theoretical basis to demonstrate the relationships between PMI and scalar product of word vectors [Arora et al., 2016].

If the word vectors satisfy Assumptions 1 and 2, and n is the number of words, then the concentration of the partition function is stated as follows [Arora et al., 2016, Lemma 2.1]:

$$\mathsf{P}[(1-\epsilon_z)\,Z \;\leq\; Z_c \;\leq\; (1+\epsilon_z)\,Z] \geq 1-\delta \tag{4}$$

for some constant Z (independent of c), $\epsilon_z = \tilde{O}(1/\sqrt{n})$ and $\delta = \exp(-\Omega(\log(n)))$. We are interested in this property since it is central for the development of all the following theorems and propositions in [Arora et al., 2016], including the relation between PMI of word pairs and the scalar product of their word vectors [Arora et al., 2016, Theorem 2.2]. Furthermore, in the experiments conducted in [Arora et al., 2016, Section 5.1], the property expressed in Equation (4) is empirically evaluated with the histogram of the partition function Z_c (which should concentrate around its mean) for word vectors obtained from common methods, such as GloVe and word2vec. By doing so, this concentration of partition function is implicitly considered as a mean to evaluate how well the word vectors follow the generative model. In this section, we will show that the property holds (modulo a small constant) not only for random vectors described by Assumptions 1 and 2, but for a set of random vectors with bounded norm and symmetrically distributed around the origin.

2.1 Preliminaries

Before presenting our main inequality, we state three lemmas whose proof are left in the appendix.

Lemma 1. *Let $\psi : \mathbb{R} \to [0, +\infty[$ be a twice continuously differentiable strictly convex even function, satisfying the following properties:*

1. $\psi'(0) = 0$
2. $\beta \mapsto \psi'(\beta)/\beta$ *is injective on* $\mathbb{R}^{+*}$
3. $\forall \beta \neq 0 \quad \psi''(0) - \psi'(\beta)/\beta > 0$
4. $\forall \beta \neq 0 \quad \psi''(\beta) - \psi'(\beta)/\beta < 0$

Then, the optimization problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{d} \psi(x_i) \triangleq \Psi(x) \\ s.t. \quad & \frac{1}{2}\|x\|^2 = \frac{1}{2}R^2 \end{aligned}$$

has the following extreme points:

1. $x^* = \pm Re_k$*, where* e_k *is the* k*-th canonical vector of* $\mathbb{R}^d$*, corresponding to global minimizers;*
2. $x_i^* = \pm\dfrac{R}{\sqrt{d}}$*, for* $i = 1, \ldots, d$*, corresponding to global maximizers.*

Proof. Cf. appendix. □

We present a similar result for the annulus domain:

Lemma 2. *Let* η *be a strictly positive real, and* $\mathbf{1}$ *the vector of ones of appropriate dimension. With the same conditions and notations as in Lemma 1, replacing the sphere of radius* R *with the annulus* Ω_η *defined by:*

$$\Omega_\eta = \{x \in \mathbb{R}^d \mid R \leq ||x||_2 \leq R + \eta\} \tag{5}$$

we have that

(i) $x = Re_k$ *is a global minimizer of* Ψ *on* Ω_η*,*
(ii) $\frac{R+\eta}{\sqrt{d}}\mathbf{1}$ *is a global maximizer of* Ψ *on* Ω_η*.*

Proof. Cf. appendix. □

Lemma 3. *Let* $L > 0$ *and consider the function* $f\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ *defined by:*

$$f(c) \quad = \quad \prod_{i=1}^{d} \begin{cases} \frac{\sinh(Lc_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ L & \text{otherwise} \end{cases} \tag{6}$$

Then $\Psi = \log(f)$ *verifies the assumptions of Lemma 1 and 2.*

Proof. Cf. appendix. □

2.2 Main Inequality

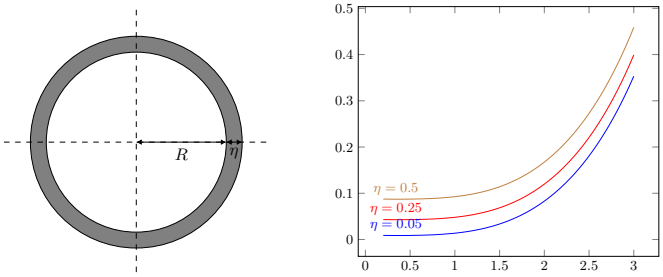
We now present our result concerning the partition function.

Proposition 1. *Let* n *be the number of words, and let us suppose the word vectors are generated independently and uniformly in a centered cube of* $\mathbb{R}^d$*. Then, if the discourse vectors belong to the annulus domain* Ω_η*, for* $R \leq 2$*, and a sufficiently small* η*, then there exists* $\gamma \ll 1$ *such that* $\forall \epsilon > 0$*, the following inequality holds with probability* $1 - \alpha$*:*

$$(1-\epsilon)(1-\gamma)\,\mathbb{E}[Z_0] \;\leq\; Z_c \;\leq\; (1+\epsilon)(1+\gamma)\,\mathbb{E}[Z_0] \tag{7}$$

where $Z_0 = Z(c_0)$*, for a constant discourse vector* c_0*, and* $\alpha \leq \exp(-\frac{1}{2}\epsilon^2 n^2)$*.*

Proof. The full proof is left in the appendix, and it is decomposed in three steps:
◦ Compute a closed form expression of the mapping $c \mapsto \mathbb{E}[Z_c]$.
◦ Study the variation of this function over Ω_η using Lemma 3.
◦ Use Berstein inequalities to bound $|Z_c - \mathbb{E}[Z_c]|$ with high probability. □



(a) Ω_η (b) Δ_η for $\eta \in \{0.05, 0.25, 0.5\}$

Fig. 1: Illustration of the maximum relative variations of $\mathbb{E}[Z_c]$ for $L = 1$ on Ω_η. (a) The annulus domain of width η. In (b), the x-axis represents the radius R considered and the y-axis the value of the maximum relative variation Δ_η (see Equation (43) in the Appendix).

Figure 1 illustrates the behavior of the maximum relative variation

$$\Delta_\eta = \frac{\max_{c \in \Omega_\eta} \mathbb{E}\left[Z_c\right] - \min_{c \in \Omega_\eta} \mathbb{E}\left[Z_c\right]}{\min_{c \in \Omega_\eta} \mathbb{E}\left[Z_c\right]}$$

for different values of η as R increases. Such behavior for small enough R and η allows us to apply the Bernstein inequality to arrive at inequality (7) with high probability.

Proposition 1 shows that the concentration property also holds for random vectors that do not necessarily satisfy Assumption 2. It is actually a fairly general property rather than an intrinsic property of word vectors and, although it was empirically verified for certain common word embedding methods in [Arora et al., 2016, Section 5.1], it does not appear as a significant quality test for word vectors. Nevertheless, this concentration property is necessary to prove the main theoretical results of [Arora et al., 2016] that we discuss in the next section.

3 Relation between PMI and scalar product

In this section, we provide an empirical evaluation of the main theorem presented in [Arora et al., 2016]. Let $p(w, w')$ be the probability of words w and w' appearing together in a window of size q in the corpus, $p(w)$ and $p(w')$ be the corresponding marginal probabilities and $v_w, v_{w'} \in \mathbb{R}^d$ the respective word vectors. Theorem 2.2 in [Arora et al., 2016] gives approximations for $\log p(w, w')$ and $\log p(w)$ as linear functions of $\|v_w + v_{w'}\|^2$ and $\|v_w\|^2$ respectively. Such approximations lead to a linear approximation of the Pointwise Mutual Information (PMI) of two words w and w':

$$\mathsf{PMI}(w, w') = \log \frac{p(w, w')}{p(w)p(w')}$$

by the scalar product $\langle v_w, v_{w'} \rangle$ of their word vectors. These results are gathered in the following theorem:

Theorem 1. *[Arora et al., 2016, Therem 2.2] Suppose the word vectors satisfy the inequality* (4)*, and the window size $q = 2$. Then*

$$\log p(w, w') = \frac{||v_w + v_{w'}||^2}{2d} - 2 \log Z \pm \epsilon, \tag{8}$$

$$\log p(w) = \frac{||v_w||^2}{2d} - \log Z \pm \epsilon, \tag{9}$$

for $\epsilon = O(\epsilon_z) + \tilde{O}(1/d) + O(\epsilon_1)$. Jointly, these imply:

$$\mathsf{PMI}(w, w') = \frac{\langle v_w, v_{w'} \rangle}{d} \pm O(\epsilon). \tag{10}$$

In the theorem, $\epsilon_z = \tilde{O}(1/\sqrt{n})$ comes from inequality (4) [Arora et al., 2016, Lemma 2.1] and ϵ_1 is from Assumption 3. See [Arora et al., 2016] for details.

For window size $q > 2$ we have the following corollary [Arora et al., 2016, Corollary 2.3].

Corollary 1. *Under the assumptions of Theorem 1, and considering $p(w, w')$ and $\mathsf{PMI}(w, w')$ for window size $q > 2$:*

$$\log p(w, w') = \frac{||v_w + v_{w'}||^2}{2d} - 2 \log Z + \Gamma \pm \epsilon, \tag{11}$$

$$\mathsf{PMI}(w, w') = \frac{\langle v_w, v_{w'} \rangle}{d} + \Gamma \pm O(\epsilon), \tag{12}$$

where $\Gamma = \log q(q-1)/2$.

In [Arora et al., 2016], it is mentioned that relation (12) is consistent with the result of [Levy and Goldberg, 2014], which showed that without dimension constraints, the solution to skip-gram with negative sampling [Mikolov et al., 2013b] corresponds to a factorization of a shifted asymmetric PMI matrix:

$$\forall w, w' \quad \mathsf{PMI}(w, w') = \langle \hat{v}_w, \hat{v}_{w'} \rangle - \beta$$

for a suitable constant β. We will discuss this result for the case of a *symmetric* PMI matrix in Section 4. In the following, we will present the results of an experimental verification of Theorem 1 and Corollary 1. Later, a discussion of these results follows.

3.1 Experimental verification

The experimental verification consists in performing a linear regression (we provide the slope, the intercept, and the coefficient of determination R^2, along with computing the Pearson correlation value) to verify the relations (8)–(12).

Word vectors Since Theorem 1 assumes that the word vectors satisfy the concentration property, we considered GloVe [Pennington et al., 2014] and SN (Squared norm) [Arora et al., 2016] word vectors because they empirically verify such property [Arora et al., 2016, Section 5.1]. We recall their respective optimization formulations.

Let $X_{w,w'}$ be the number of times words w and w' co-occur within the same window in the corpus, $f_1(X_{w,w'}) = \min(X_{w,w'}, 100)$ and $f_2(X_{w,w'}) = \min(X_{w,w'}^{3/4}, 100)$.

◦ SN formulation:

$$\min_{\{v_w\},C} \sum_{w,w'} f_1(X_{w,w'})(\log X_{w,w'} - \|v_w + v'_w\|_2^2 - C)^2$$

◦ GloVe formulation:

$$\min_{\{v_w\},\{s_w\},C} \sum_{w,w'} f_2(X_{w,w'})(\log X_{w,w'} - \langle v_w, v'_w \rangle - s_w - s'_w - C)^2$$

Note that both optimization problems are similar when $s_w = \frac{1}{2}\|v_w\|_2^2$.

Datasets All word embeddings were trained on English Wikipedia 2020. The corpus was pre-processed using the standard approach (non-textual elements removed, sentences split, tokenized). Only words appearing more than 1000 times are considered. Three different extracts from the 2020 English Wikipedia dump were used. The first corpus (denoted corpus 1) consists of the first 1 million documents deprived of prepositions and pronouns. The second corpus (denoted corpus 2) consists of the first 1,072,907 documents. The third corpus (denoted corpus 3) consists of the first 3,170,407 documents. The SN word embeddings were reproduced using code available at [Arora et al., 2018a] and GloVe word embeddings made available by [Pennington et al., 2014] were used as well.

All the results are shown in the tables 1, 2, 3. The results of tables 1 and 2 are based solely on corpus 1.

Table 1: Results for the experimental verification of equation 8 for SN word embeddings. The y-label of the linear regression is $\log p(w, w')$ and the x-label is $||v_w + v_{w'}||^2$.

Dimension	Window size	Pearson correlation	Slope	Intercept	R^2
50	2	0.73	0.0368	-23.57	0.53
100	2	0.74	0.0243	-24.75	0.55
200	2	0.76	0.0157	-26.11	0.57
300	2	0.76	0.0119	-27.00	0.58
50	10	0.78	0.0517	-27.19	0.61
100	10	0.79	0.0320	-28.40	0.62
200	10	0.79	0.0191	-29.46	0.63
300	10	0.80	0.0139	-30.02	0.64

Equations 8 and 9 It is clear that a high correlation exists between $\log p(w)$ and $||v_w||^2$, as predicted by equation 9, along with a fairly satisfying determination coefficient. However, the experimental slope of this linear relationship differs[3] slightly from the theoretical $\frac{1}{2d}$ ($= 0.01$ for $d = 50$, for example) for all of the dimensions seen in the

[3] Still, the relationship between the experimental slopes and the theoretical slopes, in an evolution w.r.t the inverse of the dimension, is satisfyingly linear both for 8 and 9.

Table 2: Results for the experimental verification of equation 9 for SN word embeddings. The y-label of the linear regression is $\log p(w)$ and the x-label is $||v_w||^2$. The partial linear regression is based on the 50 points with the highest frequencies, and its score is based on the full dataset.

Regression	Dimension	Window size	Pearson correlation	Slope	Intercept	R^2
full	50	2	0.84	0.115	-16.61	0.70
full	100	2	0.85	0.073	-18.21	0.73
full	200	2	0.89	0.044	-19.69	0.79
full	300	2	0.91	0.03	-20.29	0.83
full	50	10	0.86	0.120	-16.86	0.74
full	100	10	0.85	0.071	-17.96	0.73
full	200	10	0.87	0.040	-18.96	0.75
full	300	10	0.88	0.028	-19.41	0.78
partial	50	10	0.86	0.049	10.77	-3.58
partial	100	10	0.85	0.024	10.39	-4.69
partial	200	10	0.87	0.012	10.30	-5.17
partial	300	10	0.88	0.008	10.38	-5.16

Table 3: Results for the experimental verification of equation 10 and 12 (window size=10) for SN and GloVe word embeddings. The y-label of the linear regression is $\mathsf{PMI}(w, w')$ and the x-label is $\langle v_w, v_{w'} \rangle$. The partial linear regression is based on points with PMI less than 5.

Corpus	Embedding	Dimension	Regression	Window size	Pearson correlation	Slope	Intercept	R^2
1	SN	50	full	2	0.04	0.0061	0.93	0.002
1	SN	100	full	2	0.09	0.0088	0.89	0.008
1	SN	200	full	2	0.17	0.0109	0.86	0.03
1	SN	300	full	2	0.24	0.0117	0.85	0.06
1	SN	50	full	10	0.09	0.0129	0.32	0.008
1	SN	100	full	10	0.11	0.0094	0.29	0.011
1	SN	200	full	10	0.13	0.0070	0.27	0.02
1	SN	300	full	10	0.15	0.0063	0.27	0.02
2	SN	50	full	2	0.21	0.0360	1.07	0.04
3	SN	50	full	2	0.21	0.0378	1.14	0.04
2	SN	50	partial	2	0.17	0.0269	1.02	0.03
3	SN	50	partial	2	0.16	0.0256	1.03	0.02
2	GloVe	50	full	2	0.05	0.0082	1.17	0.003
2	GloVe	100	full	2	0.03	0.0029	1.20	0.001
3	GloVe	50	full	2	0.10	0.0179	1.21	0.011
3	GloVe	100	full	2	0.08	0.0086	1.23	0.007

experiments. For equation (9), we also performed a partial linear regression based on the 50 points with the highest frequencies. For this regression, the slope approximation from the partial linear regression is much closer to the theoretical one.

For equation 8, although the linear correlation values are satisfyingly high, the experimental slope values for windows size 2 do not match with the theoretical $\frac{1}{2d}$.

In order to empirically estimate the theoretical intercept, we approximate[4] $Z \approx 1.67 \times 10^4$, which gives $\log Z \approx 9.72$. The experimental values do not exactly match with the theoretical value of the intercept (approximately -19.44 and -9.72 for equations 8 and 9 respectively with window size 2). The error in the intercept is larger for equation 9. Perhaps the reason why the error on the intercept is smaller for equation 8 is that the SN optimization problem tries to fit equation 8.

Equation (10) For the results based on corpus 1, the correlation values are somewhat low and the determination coefficient values are poor. For the latter, the high noise of the equation (also observed in [Arora et al., 2016], see Remarks 1) is the principal reason. In fact, the experimental slope is increasing with respect to the dimension, which is completely contradictory with equation (10). Our experiments show that, as dimension goes higher, some outliers (w, w') get clustered together away from the main cluster (see figure 2). The former all verify $\frac{1}{d}\langle v_w, v_{w'}\rangle \approx 1$, while the main cluster's points have $\frac{1}{d}\langle v_w, v_{w'}\rangle \approx 0$. As a matter of fact, the marginal cluster corresponds to pairs of words occurring with themselves, because we know that for these $\frac{1}{d}\langle v_w, v_{w'}\rangle = \frac{1}{d}\,||v_w||^2 \approx 1$, when the dimension is high enough, thanks to Lemma 4. As these words tend to have medium to high PMI values, they "pull" the regression line up, hence the experimental slope increases with the dimension.

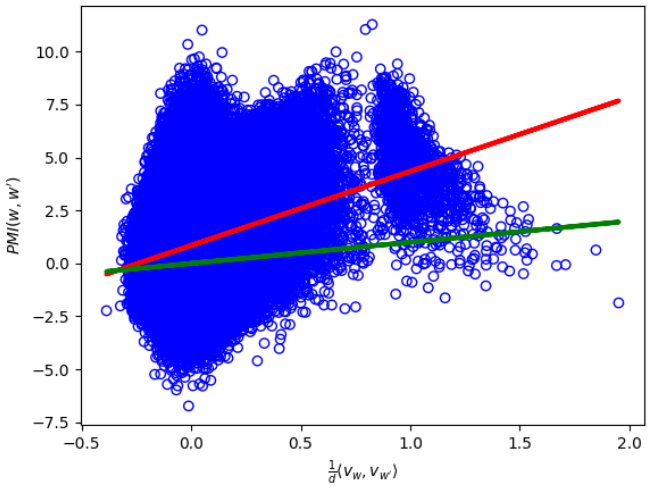


Fig. 2: Plot of PMI vs. $\frac{1}{d}\langle v_w, v_{w'}\rangle$ based on corpus 1.

[4] Z was computed as the empirical mean of sampled partition function values Z_c, computed using equation (3), by sampling random context vectors c in the unit sphere.

We then considered larger corpora and observed that results were slightly better for these, especially with regard to the slope values. The intercept values are relatively low, which is coherent with the theoretical zero value. When the window size is greater than 2, the theoretical intercept is $\gamma = \log q(q-1)/2$ according to equation (12). For windows size $q = 10$, the theoretical intercept is $\gamma \approx 3.81$. In all cases, there is a discrepancy between the experimental intercept and the theoretical one.

It remains important to visualize the shape of the plot. Indeed, figure 3 provides the plots of the experiments for different corpus sizes. Also, a heat plot is provided in order to consider the density of the plotted points. We observe that the larger the corpus, the higher the upper bound of PMI. And for this part of the plot, that is high PMI values, the linear relationship predicted by equation 10 seems nonexistent. We also observe the high discrepancy of the dot product values when $\mathsf{PMI} \approx 0$. This point will be thoroughly discussed further.

Experiments using GloVe Also, in order to give a comprehensive view of the relation between PMI and the scalar product, the relation was tested for GloVe word vectors in order to verify whether this relation's validity is an indicator of quality. From table 3, we can see that the relationship is practically nonexistent for GloVe. It is therefore possible to claim that the relation discussed is not necessary for word vectors to perform well on semantic and syntactic tasks.

3.2 Discussion

In this subsection, we discuss the relation claimed in theorem 1 between PMI and dot product of the model's word embeddings. First, we show that a distribution discrepancy exists in equation (9). Then, we provide empirical and theoretical arguments to restrict the domain where the claimed theorem can be valid. Finally, we examine granular examples of the regions of the plot to give an insight on the intrinsic difference between PMI and the scalar product.

Distribution discrepancy in equation (9) The experiments conducted to verify equation (9) show that it is not empirically verified by infrequent words. Figure 4, similarly to figure 2 in [Arora et al., 2016], shows that a linear relationship can possibly exist only when $\log p(w) > -9$ or $\frac{1}{d}\left|\left|v_w\right|\right|^2 > 1.5$. We can also provide a theoretical argument, using assumption (3) to claim that equation (9) does not hold well for infrequent words.

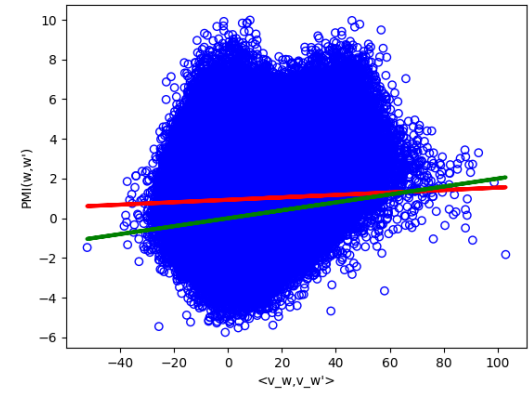
Indeed, Lemma 4 proves that $\frac{1}{d}\left|\left|v_w\right|\right|^2$ concentrates around the value 1 for a large enough value of the dimension. On the other hand, from empirical observation, logarithm word frequency seems to follow a Pareto distribution with a mode very distant from the mean of $\frac{1}{d}\left|\left|v_w\right|\right|^2$. Hence, as shown by figure 5a, there is a distribution discrepancy between $2(\log p(w) + \log Z)$ and $\frac{1}{d}\left|\left|v_w\right|\right|^2$ which strongly restricts the possible domain of validity of equation (9).

Lemma 4. *Let $X \in \mathbb{R}^d$ a real-valued random vector such that $d \in \mathbb{N}^*$.*

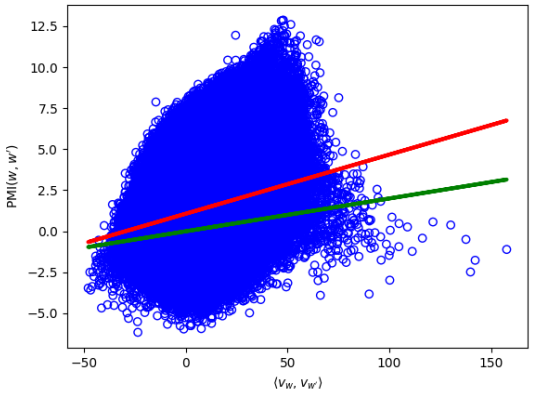
If X is drawn from the spherical Gaussian distribution in $\mathbb{R}^d$, then for all $z \in \mathbb{R}$

$$\mathbb{P}(\frac{1}{d}\left|\left|X\right|\right|^2 \geq z) = 1 - \Phi\left((z-1)\sqrt{\frac{d}{2}}\right) + \mathcal{O}(\frac{1}{\sqrt{d}}) \tag{13}$$

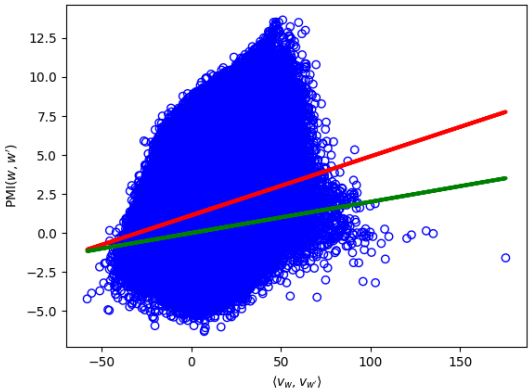
where Φ is the cumulative distribution function of the standard normal distribution.



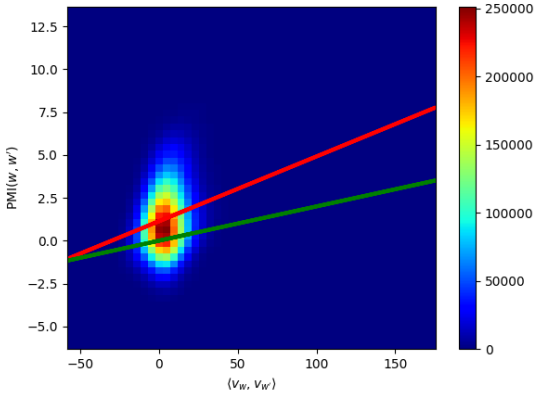
(a) Based on corpus 1



(b) Based on corpus 2



(c) Based on corpus 3



(d) Heat map of the plot based on corpus 3

Fig. 3: Experiments on equation 10. x-axis: $\langle v_w, v_{w'} \rangle$; y-axis: $\mathsf{PMI}(w, w')$. Green line: theoretical linear relationship predicted by the equation. Red line: result of the linear regression.

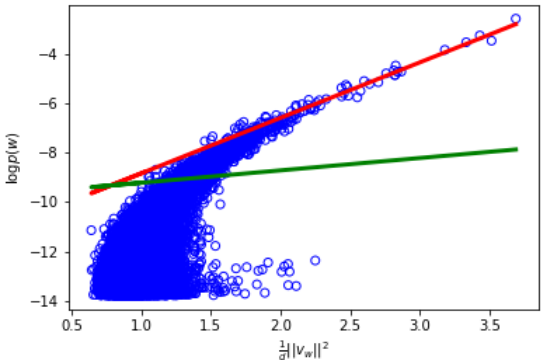
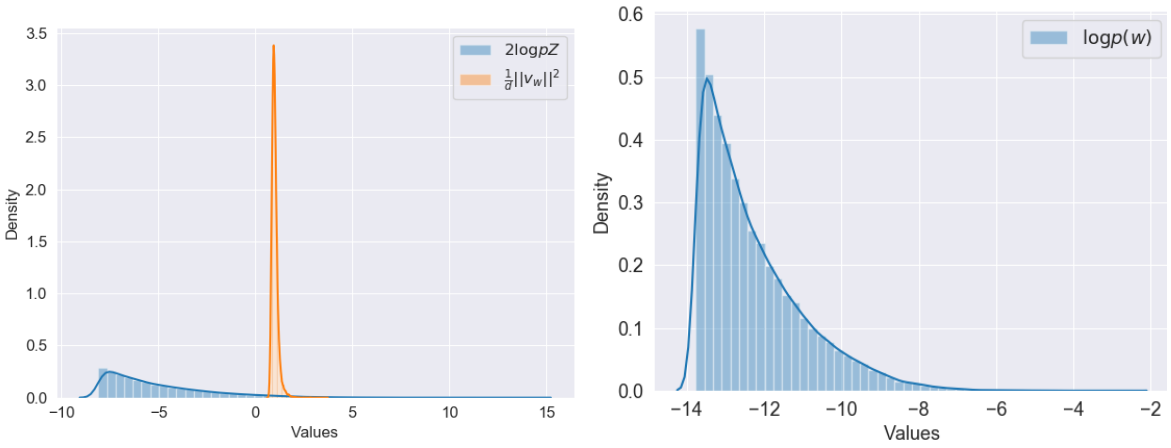


Fig. 4: Experiments on equation (9). Green line: theoretical relation predicted. Red line: result of the partial linear regression. Based on corpus 2.



(a) Density estimation for $2(\log p(w) + \log Z)$ and $\frac{1}{d}\,||v_w||^2$

(b) Density estimation for $\log p(w)$

Fig. 5: Density estimations based on corpus 2

PMI properties and restriction of theorem 1 From the results displayed on the plot 3c, we can compare the empirical upper bound of PMI and that of $\frac{1}{d}\langle\cdot,\cdot\rangle$. We empirically observe $\max \text{PMI} \approx 15$ and $\max \frac{1}{d}\langle\cdot,\cdot\rangle \approx 4$. This shows that, at least in the region where $\frac{1}{d}\langle\cdot,\cdot\rangle > 4$, we cannot have $\text{PMI} \approx \frac{1}{d}\langle\cdot,\cdot\rangle$.

In an attempt to find a restricted domain where the claimed relation is valid, we added a third dimension to the plot of PMI and $\langle\cdot,\cdot\rangle$. The result displayed in figure 6a shows that couples (w, w') for which linear relation of Theorem 1 are in general couples of infrequent words. This is not coherent with the fact that, according to [Arora et al., 2016], "very frequent words ... do not fit our model". However, this is not surprising when we consider that equations 8 and 9 seem to hold[5] better for very frequent words. In fact, in the optimization objective 3.1, very frequent pairs of words have the highest weights. Moreover, they are involved in a great number of terms, since they co-occur with a lot of words. Therefore, this can explain why the model fits better for very frequent words.

When we restrict the third dimension, that is $\max(R(w), R(w'))$, to have a threshold maximum value of 500[6], the relation seems to hold (see figure 6c).

Table 4: Words with low PMI and high scalar product

Word 1	Word 2	PMI	Scalar product	Cosine similarity
many	several	-0.82	134	0.90
march	general	0.77	80	0.55

Table 5: Words with scalar product ≈ 0 and relatively large PMI

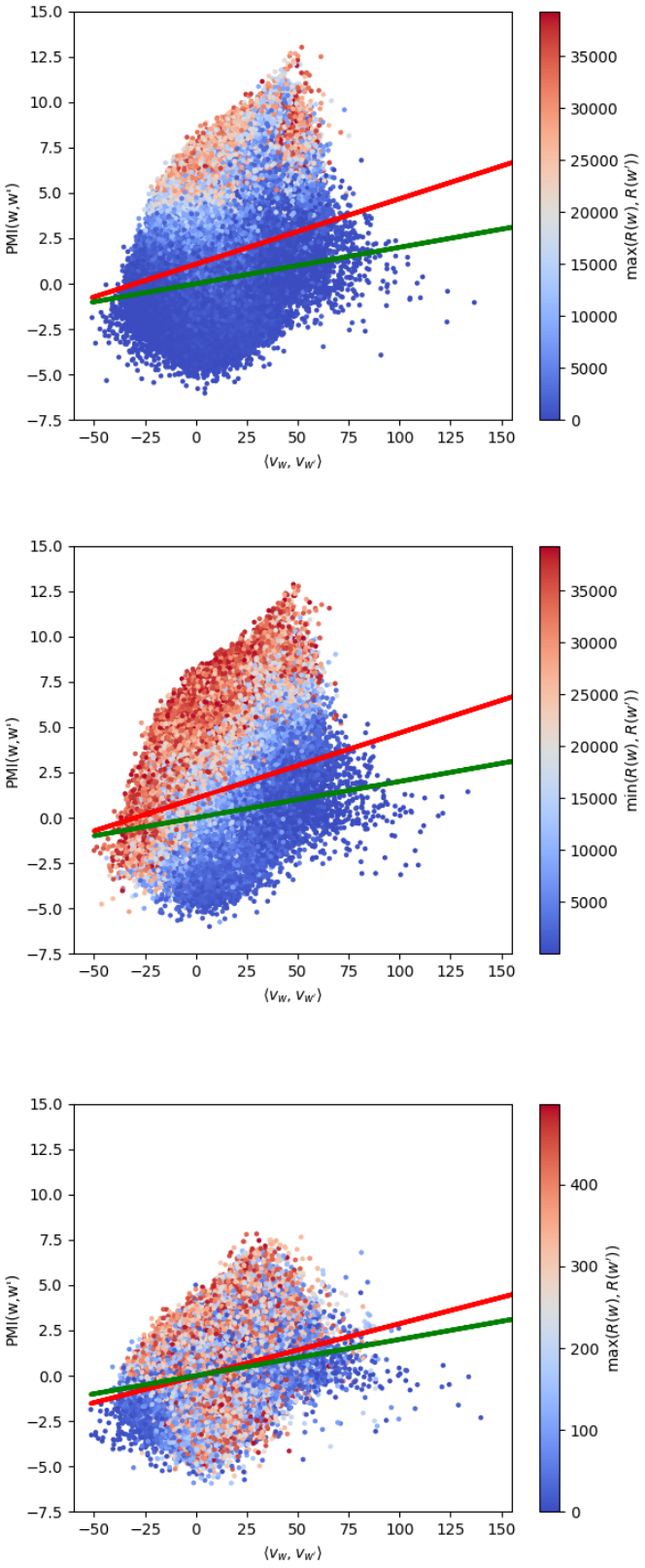
Word 1	Word 2	PMI	Scalar product	Cosine similarity
schools	newsweek	3.02	5	0.05
schools	moldovan	2.58	-5	-0.05
schools	ugandan	2.45	8	0.09

Table 6: PMI, scalar product and cosine similarity of words sampled from the corpus. The experiment was made using SN word embeddings on corpus 3.

Word 1	Word 2	PMI	Scalar product	Cosine similarity
notre	dame	10	79	0.85
obama	barack	10	81	0.92

[5] from a correlation point of view

[6] That is, a couple of word is left only if at least one of the two words of the couple is in top 500 most frequent words.



(c)

Fig. 6: Plot of $\mathsf{PMI}(w, w')$, $\langle v_w, v'_w \rangle$ with $\max(R(w), R(w'))$ in first figure, and with $\min(R(w), R(w'))$ in the second and third, where $R(w)$ is the frequency rank of word w. Based on corpus 2. Green line: theoretical relation. Red line: linear regression.

On the intrinsic difference between PMI and the dot product We advocate that PMI and the dot product, although they both encode similarity between words, do not encode the same type of similarity. While $\mathsf{PMI}(w, w')$ yields large values for words w and w' which co-occur more often than if they were independent, it merely takes into account "the company [the word] keeps" [Firth, 1957]. On the other hand, the scalar product $\langle v_w, v_{w'} \rangle$ of word embeddings trained on the non-zero entries of a global word-word co-occurrence matrix, like SN and GloVe, captures not only co-occurrences of w and w' but also those of other related words as well. Furthermore, PMI usually requires large corpora because of its unreliability with low occurrence words (see [Role and Nadif, 2011] for full details on the difficulties related to handling low occurrence events for PMI).

In order to understand the difference between PMI and the inner product of word vectors, we can distinguish three situations of couples of words: both frequent, both infrequent, one frequent and the other infrequent. Given these types, we can distinguish three regions in the plot of figure 6a. It can be inferred from two first plots of figure 6 that the top region of the surface corresponds to infrequent words. The bottom right region corresponds to frequent words. The bottom left corresponds to couples made of a frequent and an infrequent word.

When both words are infrequent and happen to co-occur, it is very likely that they have high PMI and scalar product values. This explains the shape of the top region of the surface on the plot. This part of the plot is the most interesting as it is where the major outliers of theorem 1 live. These values always exist in natural language. To illustrate this, table 6 contains an example of very high values of PMI. Usually, these words would rarely appear without their partner word, thus the high PMI value. The scalar product and cosine similarity are also high which is coherent for such words that rarely appear without the other.

Table 4 contains a sample of words from the region of high discrepancy around the 0 PMI value in figure 3. This is an example of very similar words, as inferred by the cosine similarity and scalar product, with low PMI values. Especially for the words 'many' and 'several' which are similar but will never appear together[7], thus the low PMI value. The scalar product was able to capture the similarity because it had access to all the contexts of 'many' and 'several' and inferred that these were similar.

In table 5 we can see a sample of words with scalar product ≈ 0 and positive[8] PMI value. We can give the following explanation for 'schools' and 'moldovan' for example: 'moldovan' is a relatively rare word and it happens that it naturally occurs often (relatively to the frequency of 'moldovan') with 'schools', thus the PMI value. But these words are completely different semantically, thus the scalar product value. Here is another example of unwanted behavior for low occurrence words.

Finally, an important difference between the scalar product and PMI can be observed for words occurring with themselves. For this type of co-occurrences, the scalar products are naturally very large[9]. However, the PMI values can be anywhere from negative to large positive values: words 'the' and 'her' have a dot product of 98 and PMI value of -1.26, while 'as' and 'well' have a dot product of 136 and PMI value of 4.58. In fact, this last example is the exact type of co-occurrences causing the bottom-right edge on the scatter plot (see subfigure (c) of figure 3). This further justifies how a strict linear relationship between PMI and scalar product can hardly exist.

4 Relation with implicit matrix factorization

The experimental results of Section 3 point in the direction that Equation (12) is not verified in practice with a small noise level ϵ. In this section we shall prove that, as long as a symmetric PMI matrix is considered, (12) cannot

[7] Usually, redundancy is avoided in writings and such interchangeable words would not be used together.

[8] These values are relatively large as it is useful to notice that when $\mathsf{PMI}(w, w') \approx 3$, words w and w' are 20 times more likely to co-occur than if they were independent.

[9] being greater than 500 is considered very large

hold if the noise ϵ vanishes. To this end, we will show that the shifted symmetric PMI matrix fails to be positive semidefinite when considering natural language.

In [Levy and Goldberg, 2014], it was shown that the skip-gram with negative sampling [Mikolov et al., 2013b] corresponds to an implicit matrix factorization:

$$\forall w, c, \quad \langle v_w, v_c \rangle = \mathsf{PMI}(w, c) - \beta, \tag{14}$$

where $\beta = \log k$, k is the number of "negative samples" and $\mathsf{PMI}(w, c)$ corresponds to an entry of an asymmetric (usually rectangular) word-context PMI matrix. Each context is defined by a window of size q around each token w_ℓ, i.e $w_{\ell-q}, \ldots, w_{\ell-1}, w_{\ell+1}, \ldots, w_{\ell+q}$ is the context for the word/token w_ℓ. In (14), $v_w, v_c \in \mathbb{R}^d$ for a suitable dimension d.

If we consider a matrix V whose rows are the vectors v_w, and C a matrix whose rows are the vectors v_c, then (14) can be written in matrix form as

$$VC^\top = M - \beta \mathbf{1}_{|V|} \mathbf{1}_{|C|}^\top, \tag{15}$$

where M is a $|V| \times |C|$ matrix with entries $M_{wc} = \mathsf{PMI}(w, c)$ and $\mathbf{1}_m$ denotes the vector of ones in $\mathbb{R}^m$. The singular value decomposition [Golub and Van Loan, 1996] ensures that (15) holds for some $d = \text{rank}(VC^\top) \leq \text{rank}(M) + 1$.

In view of relation (12), one may wonder whether (14) also holds for a *symmetric* PMI matrix: here the vocabularies of words and contexts are the same.

In fact, in [Arora et al., 2016, pg. 389] one finds: "This [Equation (12) in Corollary 1] is also consistent with the shift β for fitting PMI in (Levy and Goldberg, 2014b), which showed that without dimension constraints, the solution of skip-gram with negative sampling satisfies $\mathsf{PMI}(w, w') - \beta = \langle v_w, v_{w'} \rangle$ for a constant β that is related to the negative sampling in the optimization. Our result justifies via a generative model why this should be satisfied even for low dimensional word vectors."

Let us assume that there exists a scalar β such that

$$\forall w, w', \quad \langle v_w, v_{w'} \rangle = \mathsf{PMI}(w, w') - \beta. \tag{16}$$

Suppose the vocabulary is finite (of size n), and since $\mathsf{PMI}(w, w') = \mathsf{PMI}(w', w)$, define the symmetric matrix M, such that $M_{w,w'} = \mathsf{PMI}(w, w')$. Then, we can write (16) in matrix form as

$$VV^\top = M - \beta \mathbf{1}\mathbf{1}^\top,$$

where V is a $n \times d$ matrix whose rows contain the vectors $v_w \in \mathbb{R}^d$, and $\mathbf{1} \in \mathbb{R}^n$ denotes the vector of ones.

Since $VV^\top$ is symmetric positive semidefinite, we obtain that, for every vector $y \in \mathbb{R}^n$

$$0 \leq y^\top (M - \beta \mathbf{1}\mathbf{1}^\top) y = y^\top M y - \beta (\mathbf{1}^\top y)^2.$$

In particular, taking $y \in \{\mathbf{1}\}^\perp$, we have

$$\forall y \in \{\mathbf{1}\}^\perp, \quad y^\top M y \geq 0. \tag{17}$$

Let w and w' be a pair of words for which $p(w,w') > 0$, $p(w,w) > 0$, $p(w',w') > 0$, and choose $y = e_w - e_{w'} \in \{\mathbf{1}\}^{\perp}$, where $e_w, e_{w'}$ are canonical vectors of $\mathbb{R}^n$. Thus,

$$\begin{aligned} y^\top M y =& (e_w - e_{w'})^\top M (e_w - e_{w'}) \\ =& M_{ww} - 2M_{ww'} + M_{w'w'} \\ =& \mathsf{PMI}(w,w) - 2\mathsf{PMI}(w,w') \\ &+ \mathsf{PMI}(w',w') \\ =& \log \frac{p(w,w)}{p(w)p(w)} - 2\log \frac{p(w,w')}{p(w)p(w')} \\ &+ \log \frac{p(w',w')}{p(w')p(w')} \\ =& \log p(w,w) - 2\log p(w,w') \\ &+ \log p(w',w'). \end{aligned}$$

The last inequality leads to

$$\log \frac{p(w,w)p(w',w')}{p(w,w')^2} \geq 0$$

or, equivalently,

$$p(w,w')^2 \leq p(w,w)p(w',w'). \tag{18}$$

However, this inequality is violated by a pair of words w and w' for which $p(w,w)$ and $p(w',w')$ are quite small when compared to $p(w,w')$, i.e words that appear repeated in very few windows but co-occur considerably more as illustrated in the following examples based on the statistics for the "corpus 2":

- **Example 1:** If $w =$ professional and $w' =$ wrestler. Then, $p(w,w') = 4.51 \times 10^{-6}$ and $p(w,w) = 2.09 \times 10^{-7}$ and $p(w',w') = 5.26 \times 10^{-8}$. In this case $p(w,w')^2 > p(w,w)p(w',w')$.
- **Example 2:** If we consider w, w' as the pair of words *well*, *done* (respec.), we have $\log p(w,w) \approx -14.7547$, $\log p(w',w') \approx -17.5806$ and $\log p(w,w') \approx -13.9783$, which shows that $2\log p(w,w') > \log p(w,w) + \log p(w',w')$, i.e inequality (18) does not hold.

Hence, condition (18), which is a necessary condition for (16), can be violated with natural language, thereby invalidating the claim (16), regardless the dimension d and the constant β. Therefore, Equation (12) with zero noise/error does not hold.

5 Conclusion

The empirical verification of the equations listed by Theorem 1 and Corollary 1 strongly suggests that the claimed linear relation between PMI and the inner product of word embeddings does not hold in practice – even for word vectors satisfying the concentration property (Equation 4) – unless an unacceptably high error term $O(\epsilon)$ is tolerated. Moreover, the statistical discussion in section 3.2 provides evidence of the existence of a range of values where the linear relation cannot hold.

These experimental findings concerning the violation of Equation (10) (and Equation (12)) – with error terms dropped – are further corroborated by the theoretical analysis of Section 4 which shown that the desired linear relation $\langle v_w, v_{w'} \rangle = \mathsf{PMI}(w,w') - \beta$ implies in the positiveness of the symmetric PMI matrix in a certain subspace, but such condition can be violated by natural language.

Section 2 showed that the concentration of partition function is rather a general property and the aforementioned arguments go against the linear relation between PMI and scalar product. Therefore, we advocate that neither should be considered as a quality test for word embeddings.

Furthermore, the failure of word vectors verifying the concentration property (empirically) to satisfy Equation (12) leads to the deduction that either natural language does not follow the assumptions of the generative model (i.e. word vectors and discourse vectors fulfilling Assumptions 1 and 3 cannot co-exist) or the current methods for word embeddings do not allow us to reconstruct word vectors aligned with the model.

References

Arora et al., 2018a. Arora, S., Khodak, M., Saunshi, N., and Vodrahalli, K. (2018a). A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and lstms. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Arora et al., 2016. Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016). A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Arora et al., 2018b. Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2018b). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Devlin et al., 2019. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Firth, 1957. Firth, J. R. (1957). A synopsis of linguistic theory.

Golub and Van Loan, 1996. Golub, G. A. and Van Loan, C. (1996). *Matrix Computations, 3rd edition*. The John Hopkins University Press, London.

Hochreiter and Schmidhuber, 1997. Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Khalife et al., 2019. Khalife, S., Liberti, L., and Vazirgiannis, M. (2019). Geometry and analogies: a study and propagation method for word representations. In *International Conference on Statistical Language and Speech Processing*, pages 100–111. Springer.

Levy and Goldberg, 2014. Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Mikolov et al., 2013a. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov et al., 2013b. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pennington et al., 2014. Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Role and Nadif, 2011. Role, F. and Nadif, M. (2011). Handling the impact of low frequency events on co-occurrence based measures of word similarity.

Vaswani et al., 2017. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Appendix

Lemma 1. *Let $\psi : \mathbb{R} \to [0, +\infty[$ be a twice continuously differentiable strictly convex even function, satisfying the following properties:*

1. $\psi'(0) = 0$
2. $\beta \mapsto \psi'(\beta)/\beta$ *is injective on* $\mathbb{R}^{+*}$
3. $\forall \beta \neq 0 \quad \psi''(0) - \psi'(\beta)/\beta > 0$
4. $\forall \beta \neq 0 \quad \psi''(\beta) - \psi'(\beta)/\beta < 0$

Then, the optimization problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{d} \psi(x_i) \triangleq \Psi(x) \\ s.t. \quad & \frac{1}{2}\|x\|^2 = \frac{1}{2}R^2 \end{aligned}$$

has the following extreme points:

1. $x^* = \pm R e_k$*, where* e_k *is the* k*-th canonical vector of* $\mathbb{R}^d$*, corresponding to global minimizers;*
2. $x_i^* = \pm \dfrac{R}{\sqrt{d}}$*, for* $i = 1, \ldots, d$*, corresponding to global maximizers.*

Proof. Let us consider the first order optimality conditions. The Lagrange equations are

$$\nabla \Phi(x) + \lambda x = 0$$

or equivalently

$$\forall i \in \{1, \ldots, d\} \qquad \psi'(x_i) + \lambda x_i = 0 \tag{19}$$

For the remaining of this proof, let $(x, \lambda) \in \mathbb{R}^d \times \mathbb{R}$ be a fixed vector and scalar verifying Equation (19). Such x and λ exists because we are considering a continuous function over a compact set, thus it attains a maximum and a minimum in the feasible set. Notice that $x_i = 0$ solves this equation for any λ. However, we cannot set $x_i = 0$ for every $i \in \{1, \ldots, d\}$, because $x = 0$ is infeasible.

Therefore, there should be components some components of x verifying $x_i \neq 0$. For the non-zero components of x, Equation (19) must hold for the same λ. Since the gradient of the constraint does not vanish at any feasible point, the Linear Independence Constraint Qualification (LICQ) holds and hence there exists λ fulfilling Equation (19) for some feasible point.

First, we remark that $\lambda \neq 0$. Indeed, if $\lambda = 0$, then from Equation (19), $\forall i \ f'(x_i) = 0$, but f is convex and $f'(0) = 0$, which implies that $x_i = 0$ for all i, leading to an infeasible point.

Thus, for the non-zero components of x, from Equation (19), we obtain

$$x_i \neq 0 \implies \lambda = -\frac{\psi'(x_i)}{x_i} \neq 0$$

But, since $\beta \mapsto \psi'(\beta)/\beta$ is injective on $\mathbb{R}^{+*}$, we conclude that the non-zero components of x must be all equal, i.e $\exists \beta^* > 0$ s.t. $\forall i \ x_i \neq 0 \implies x_i = \beta^*$. From the feasibility of x, we conclude that

$$\beta^* = \pm \frac{R}{\sqrt{\|x\|_0}}$$

where $\|x\|_0$ denotes the number of non-zero entries of x.

Let us now analyze the second order conditions for the feasible points verifying Equation (19). Since the objective function is separable, the Hessian of the Lagrangian $\nabla^2_{xx}L(x,\lambda)$ is a diagonal matrix whose diagonal entries verify $\forall i \in \{1,\dots,d\}$:

$$[\nabla^2_{xx}L(x,\lambda)]_{ii} = \psi''(x_i) - \psi'(\beta^*)/\beta^*$$

From the properties of f, we can deduce

$$[\nabla^2_{xx}L(x,\lambda)]_{ii} = \begin{cases} \psi''(0) - \frac{f'(\beta^*)}{\beta^*} & \text{if } x_i = 0 \\ \psi''(\beta^*) - \frac{\psi'(\beta^*)}{\beta^*} & \text{otherwise} \end{cases}$$

For the remaining of this proof, for given α and β, let $\delta(\alpha,\beta) \triangleq \psi''(\alpha) - \psi'(\beta)/\beta$. We remind our reader that, by assumption, $\beta \neq 0 \implies \delta(0,\beta) > 0$ and $\delta(\beta,\beta) < 0$.

Therefore, for a given $y \in \mathbb{R}^d$, we have

$$\begin{aligned} y^T\nabla^2_{xx}L(x,\lambda)y = \delta(0,\beta^*) \sum_{i:x_i=0} y_i^2 \\ + \delta(\beta^*,\beta^*) \sum_{i:x_i\neq 0} y_i^2 \end{aligned}$$

If all components of x are non-zero, then we get $\forall y \in \mathbb{R}^d \setminus \{0\}$:

$$y^T\nabla^2_{xx}L(x,\lambda)y = \delta(\beta^*,\beta^*) \sum_{i:x_i\neq 0} y_i^2 < 0$$

Also, we already proved non zero components of x must be equal; this proves that x verifying $\forall i \in \{1,\dots,d\}, \;\; x_i = \pm\frac{R}{\sqrt{d}}$ satisfy the second order sufficient conditions for a local maximizer.

Now, let us show that if x has at least one zero component and more than one non-zero components, then x is a saddle-point. Without loss of generality, assume that exactly two entries of x are non-zero, then due to the previous discussion, they must be equal, e.g. $x^T = (0,\dots,0,\beta,\beta)$. The sufficient second order conditions concern the Hessian of the Lagrangian with respect to primal variables, which should be positive definite when restricted on the linear null space of the Jacobian of the constraint inequalities. In this case, this linear space is given by:

$$\begin{aligned} x^\perp = \{y \in \mathbb{R}^d \; : \; y = & (w_1,\dots,w_{d-2},\alpha,-\alpha) \\ & , w \in \mathbb{R}^{d-2}, \alpha \in \mathbb{R}\} \end{aligned}$$

In particular, choosing

$$y = (w_1,0,\dots,0,\alpha,-\alpha) \in x^\perp$$

we obtain

$$y^T\nabla^2_{xx}L(x,\lambda)y = \delta(0,\beta^*)w_1 + 2\delta(\beta^*,\beta^*)\alpha^2$$

Then:

i) $w_1 > 0 \quad \alpha = 0 \implies y^T\nabla^2_{xx}L(x,\lambda)y > 0$
ii) $w_1 = 0 \quad \alpha \neq 0 \implies y^T\nabla^2_{xx}L(x,\lambda)y < 0$

This implies that x is neither a minimizer nor a maximizer.
Finally, if $x = \pm Re_k$, for some canonical vector e_k, we obtain, for every $y \in x^\perp \setminus \{0\}$,

$$\begin{aligned} y^T \nabla^2_{xx} L(x, \lambda) y &= \delta(0, \beta^*) \sum_{i:x_i=0} y_i^2 + \delta(\beta^*, \beta^*) \times 0 \\ &= \delta(0, \beta^*) \sum_{i:x_i=0} y_i^2 > 0 \end{aligned}$$

which proves that $x = \pm Re_k$ satisfies the second order sufficient conditions for a local minimizer.

Furthermore, since f is even, and the maximizers (and minimizers) described above only differ by the sign of their entries, we can conclude that all of them are global.

□

Lemma 2. *Let η be a strictly positive real, and $\mathbf{1}$ the vector of ones of appropriate dimension. With the same conditions and notations as in Lemma 1, replacing the sphere of radius R with the annulus Ω_η defined by:*

$$\Omega_\eta = \{x \in \mathbb{R}^d \mid R \leq ||x||_2 \leq R + \eta\} \tag{5}$$

we have that

(i) $x = Re_k$ is a global minimizer of Ψ on Ω_η,
(ii) $\frac{R+\eta}{\sqrt{d}}\mathbf{1}$ is a global maximizer of Ψ on Ω_η.

Proof. Both (i) and (ii) can be proved in two steps:

(i) Since ψ is even, we limit the study on the set of positive vectors. We show that the maximum of ψ is reached on the sphere of radius $R + \eta$, and on the sphere of radius R for the minimum. This can be proved by remarking that:

$$\begin{aligned} x > 0, \quad x \in \mathring{\Omega}_\eta \quad \text{and} \quad R < \lambda ||x|| < R + \eta \\ \implies \lambda x \in \Omega_\eta \quad \text{and} \quad \psi(\lambda x) > \psi(x) \end{aligned}$$

Which can be deduced by the fact that ψ is strictly convex and $\psi'(0) = 0$, hence ψ is increasing on $\mathbb{R}^+$. This implies that the minimum of Ψ is reached on the sphere of radius R, and its maximum on the sphere of radius $R + \eta$.

(ii) Then, we use Lemma 1 to conclude. □

Lemma 3. *Let $L > 0$ and consider the function f: $\mathbb{R}^d \longrightarrow \mathbb{R}$ defined by:*

$$f(c) \quad = \quad \prod_{i=1}^{d} \begin{cases} \frac{\sinh(Lc_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ L & \text{otherwise} \end{cases} \tag{6}$$

Then $\Psi = \log(f)$ verifies the assumptions of Lemma 1 and 2.

Proof. In order to simplify the expressions, we will consider that $L = 1$ but the general case can be treated similarly. First, let us consider the function

$$\phi : x \mapsto \begin{cases} \frac{\sinh(c_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

And in the following, let $\psi = \log \phi$.

i) First, ψ is twice continuously differentiable. Indeed, ψ is continous on $\mathbb{R}$ and

$$\lim_{x \to 0} \psi'(x) = 0 \tag{20}$$

So with the derivation extension theorem, ψ is differentiable in 0 and $\psi'(0) = 0$. We use the same reasoning with ϕ' and show that ψ is twice differentiable on $\mathbb{R}$, $\psi''(0) = \frac{1}{3}$.

ii) ψ strictly logarithmically convex by composition since:
- $\log$ is strictly increasing on $\mathbb{R}^{+*}$
- ϕ is strictly convex on $\mathbb{R}$, this can be seen from its second derivative:

$$\forall x \neq 0 \quad \phi''(x) = \frac{-1 - 2x^2 + \cosh(2x)}{2x^4} > 0$$

which can be deduce from the Taylor series of $\cosh$.

iii) ψ is even since ϕ is. Besides, as proved in i), we have $\psi'(0) = 0$ and $\psi''(0) = \frac{1}{3}$. Furthermore, for $x \neq 0$:

$$\begin{aligned} \psi''(0) - \frac{\psi'(x)}{x} &= \frac{1}{3} - \frac{\psi'(x)}{x} \\ &= \frac{1}{3} - \frac{1}{\sinh(x)}\left(\frac{\cosh(x)}{x} - \frac{\sinh(x)}{x^2}\right) \end{aligned}$$

and

$$\begin{aligned} \psi''(x) - \frac{\psi'(x)}{x} = &-x\frac{\coth(x)}{\sinh(x)}\left(\frac{\cosh(x)}{x} - \frac{\sinh(x)}{x^2}\right) \\ &+\frac{x}{\sinh(x)}\left(-2\frac{\cosh(x)}{x^2} + 2\frac{\sinh(x)}{x^3} + \frac{\sinh(x)}{x}\right) \end{aligned}$$

Now, let us prove:

iv) $\forall x \neq 0, \quad \psi''(0) - \frac{\psi'(x)}{x} > 0$
v) $\forall x \neq 0, \quad \psi''(x) - \frac{\psi'(x)}{x} < 0$
vi) $x \mapsto \frac{\psi'(x)}{x}$ is injective on $\mathbb{R}^{+*}$.

After computations, we remind that:

$$\phi'(x) = \frac{x\cosh x - \sinh x}{x^2}$$

$$\phi''(x) = \frac{(x^3 + 2x)\sinh x - 2x^2\cosh x}{x^4}$$

In particular:

$$\psi'(x) = \frac{\phi'(x)}{\phi(x)} \tag{21}$$

$$= \frac{x\cosh x - \sinh x}{x\sinh x} \tag{22}$$

$$= \frac{\cosh x}{\sinh x} - \frac{1}{x} = \coth x - \frac{1}{x}$$

$$= \left(\frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \cdots\right) - \frac{1}{x} \tag{23}$$

$$= \frac{x}{3} - \frac{x^3}{45} + \cdots \tag{24}$$

Now, let us consider the function q be defined as:

$$q(x) = \begin{cases} \dfrac{\psi'(x)}{x} & x \neq 0 \\ q(0) = \frac{1}{3} & \text{otherwise} \end{cases}$$

After some algebraic manipulation and Taylor series expansion of $\coth$, we obtain

$$\forall x \neq 0 \qquad q(x) = \frac{-1 + x\coth x}{x^2} \tag{25}$$

$$= \frac{-1 + x\left(\frac{1}{x} + \frac{x}{3} + \cdots\right)}{x^2} \tag{26}$$

$$= \frac{1}{3} - \frac{x^2}{45} + 2\frac{x^4}{945} + \cdots \tag{27}$$

$$q'(x) = \frac{2 - x(\coth x + x\,\mathrm{csch}^2 x)}{x^3} \tag{28}$$

$$= \frac{1}{15}(\frac{1}{3} - 1)x + \frac{1}{189}(\frac{1}{5} - 1)2x^3 + \ldots \tag{29}$$

which implies: $\forall x > 0 \quad \frac{\psi'(x)}{x} = q'(x) < 0$. This proves that the function q is injective on $\mathbb{R}^{+*}$ (q is strictly decreasing because q' is strictly negative). Property vi) is proved.

Besides,

$$q''(x) = \frac{1}{15}(\frac{1}{3} - 1) + \frac{1}{189}(\frac{1}{5} - 1)6x^2 + \ldots \quad < 0$$

hence $q'(0) = 0$ and $\forall x \quad q''(x) < 0$, implying that $q(0) = \frac{1}{3}$ is the global maximum: $\forall x \in \mathbb{R} \quad q(x) \leq 1/3$.

Moreover,

$$\begin{aligned}\psi''(x) &= \frac{\phi''(x)\phi(x) - \phi'(x)^2}{\phi(x)^2} = \frac{1}{x^2} - \operatorname{csch}^2 x \\ &= \frac{1}{x^2} - \left(\frac{1}{x^2} - \frac{1}{3} + \frac{x^2}{15} - \cdots\right) \\ &= \frac{1}{3} - \frac{x^2}{15} + \cdots\end{aligned}$$

implying

$$\forall x \neq 0 \quad \psi''(0) - \frac{\psi'(x)}{x} = \frac{1}{3} - q(x) > 0$$

showing Property iv).

Finally, for $x \neq 0$:

$$\begin{aligned}\psi''(x) - \frac{\psi'(x)}{x} &= \frac{2 - x(\coth x + x \operatorname{csch}^2 x)}{x^2} \\ = \frac{1}{15}(\frac{1}{3} - 1)x^2 + \frac{1}{189}(\frac{1}{5} - 1)2x^4 + \cdots < 0\end{aligned}$$

proving v).

□

Proposition 1. *Let n be the number of words, and let us suppose the word vectors are generated independently and uniformly in a centered cube of $\mathbb{R}^d$. Then, if the discourse vectors belong to the annulus domain Ω_η, for $R \leq 2$, and a sufficiently small η, then there exists $\gamma \ll 1$ such that $\forall \epsilon > 0$, the following inequality holds with probability $1 - \alpha$:*

$$(1-\epsilon)(1-\gamma)\,\mathbb{E}[Z_0] \;\leq\; Z_c \;\leq\; (1+\epsilon)(1+\gamma)\,\mathbb{E}[Z_0] \tag{7}$$

where $Z_0 = Z(c_0)$, for a constant discourse vector c_0, and $\alpha \leq \exp(-\frac{1}{2}\epsilon^2 n^2)$.

Proof. Let $v, c \in \mathbb{R}^d$ be the word and discourse vectors, respectively, with the following properties:

$$\|v\| \leq \kappa \tag{30}$$

$$\mathbb{E}\left[\langle v, c\rangle\right] = 0 \tag{31}$$

From (30) and Cauchy-Schwarz inequality

$$\langle v, c\rangle \leq |\langle v, c\rangle| \leq \|v\|\|c\| \leq 3\kappa \tag{32}$$

where we suppose $||c|| \leq 3$ by assumption. It follows that

$$\exp\langle v, c\rangle \leq \exp 3\kappa \tag{33}$$

Since the random vectors v are i.i.d. and by convexity of the exponential, we have from (31)

$$\begin{aligned}\mathbb{E}\left[Z_c\right] = n\mathbb{E}\left[\exp\langle v, c\rangle\right] &\geq n \exp \mathbb{E}\left[\langle v, c\rangle\right] \\ &\geq n \exp(0) = n\end{aligned} \tag{34}$$

Moreover, we are also able to bound the variance of Z_c:

$$\begin{aligned}\text{Var}\left[Z_c\right] &= \sum_v \text{Var}\left[\exp\langle v, c\rangle\right] = n\text{Var}\left[\exp\langle v, c\rangle\right] \\ &\leq n\mathbb{E}\left[\exp 2\langle v, c\rangle\right] \\ &\leq n\mathbb{E}\left[\exp(6\kappa)\right] = \exp(6\kappa)n\end{aligned}$$

Now let Λ be the constant defined as follows:

$$\Lambda = \exp(6\kappa)$$

Let $\epsilon > 0$. Thanks to (33) and (5), we can apply the Bernstein's inequality to the sum of random variables $Z_c = \sum_v \exp\langle v, c\rangle$, to obtain

$$P[|Z_c - \mathbb{E}\left[Z_c\right]| > \varepsilon n] \leq \exp\left(-\frac{\frac{1}{2}\varepsilon^2 n^2}{n\Lambda + \frac{1}{3}\sqrt{\Lambda}\varepsilon n}\right) \tag{35}$$

and from (34)

$$P[|Z_c - \mathbb{E}\left[Z_c\right]| > \varepsilon\mathbb{E}\left[Z_c\right]] \leq \exp\left(-\frac{\frac{1}{2}\varepsilon^2 n^2}{n\Lambda + \frac{1}{3}\sqrt{\Lambda}\varepsilon n}\right) \tag{36}$$

which shows the concentration of Z_c around $\mathbb{E}\left[Z_c\right]$ for *any* fixed unit norm vector c.

Let us show now that $\mathbb{E}\left[Z_c\right]$ does not vary much with c. To this end, we need additional assumptions about the distribution of v apart from (30) and (31). We are interested in $\mathbb{E}[Z_c]$, and in particular the amplitude of its variation with respect to c. If the word vectors admit a density function ξ, then:

$$\mathbb{E}_v[exp(\langle v, c\rangle)] = \int_\Omega exp(\langle v, c\rangle)\xi(v)dv \tag{37}$$

If the word vectors are independent and identically distributed, it should be noted that:

$$\mathbb{E}_v[Z_c] = n\,\mathbb{E}_v[exp(\langle v, c\rangle)] \tag{38}$$

where n is the number of words. Firstly, in order to simplify the calculation, we will consider that v is distributed uniformly on Ω which is the cube of $\mathbb{R}^d$ centered in 0, of side length $2L$. Then, integration using Fubini Theorem yields:

$$\mathbb{E}_v[exp(\langle v, c\rangle)] = 2^d \prod_{i=1}^{d} \frac{\sinh(Lc_i)}{c_i} \tag{39}$$

Consider the function $f\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by:

$$f(c) \quad = \quad \prod_{i=1}^{d} \begin{cases} \frac{\sinh(Lc_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ L & \text{otherwise} \end{cases} \tag{40}$$

We will first discuss the variations in the amplitude of f on the sphere $\mathcal{S}_R$ centered in 0 with radius R. The relative amplitude of the variations of f on $\mathcal{S}_R$ is given by:

$$\frac{\max_{c\in\mathcal{S}_R} f(c) - \min_{c\in\mathcal{S}_R} f(c)}{\min_{c\in\mathcal{S}_R} f(c)} \tag{41}$$

where $f(x) = \mathbb{E}[Z_c]$.

Using Lemmas 3 and 1, we can infer the two following properties:

- On the one hand, f reaches its maximum at a point c such that $c_1 = c_2 = \ldots = c_d = \frac{R}{\sqrt{d}}$. And then

$$\max_{c \in S_R} f(c) = [\frac{\sqrt{d}}{R} \sinh(\frac{LR}{\sqrt{d}})]^d \tag{42}$$

- On the other hand, the minimum of f is reached for a point where every coordinate has been set to 0 except one (such point exists on the sphere), and therefore, f reaches its minimum on a point c such that

$$\phi(c_1) = \ldots = \phi(c_{d-1}) = L$$
$$\text{and} \quad \phi(c_d) = \frac{\sinh(LR)}{R}$$

Hence,

$$\min_{c \in \mathcal{S}_R} f(c) = L^{d-1} \frac{\sinh(LR)}{R}$$

A first interesting result is that the extrema of f do not depend on the dimension if $L = 1$.

It should be noted that the absolute variations of $\mathbb{E}[Z_c] = n\, 2^d\, f(c)$ increases exponentially with respect to the dimension d and linearly with respect to the number of words n, the maximum ***relative variation*** of $\mathbb{E}[Z_c]$ in Equation (41) is the same as f.

Now, let us observe the behavior of the maximum of f, when the dimension d tends to infinity. The Taylor expansion at order 3 of $\sinh$ in 0 is given by:

$$\sinh(x) = x + \frac{x^3}{6} + o(x^3)$$

Therefore, using properties of the exponential:

$$\begin{aligned}
\max_{c \in \mathcal{S}_R} f(c) &\underset{d \to +\infty}{=} (\frac{\sqrt{d}}{R})^d [\, \frac{LR}{\sqrt{d}} + \frac{1}{6}(\frac{LR}{\sqrt{d}})^3 + o(\frac{LR}{\sqrt{d}})^3 \,]^d \\
&= (L + \frac{LR^2}{6d} + o(\frac{1}{d}))^d \\
&= L^d (1 + \frac{R^2}{6d} + o(\frac{1}{d}))^d \\
&\underset{d \to +\infty}{\sim} L^d e^{\frac{R^2}{6}}
\end{aligned}$$

Then, if $d \gg 1$ (e.g $d \geq 50$):

$$\begin{aligned}
\Delta(R) &= \frac{\max_{c \in \mathcal{S}_R} f(c) - \min_{c \in \mathcal{S}_R} f(c)}{\min_{c \in \mathcal{S}_R} f(c)} \\
&\underset{d \to +\infty}{\sim} L \frac{e^{\frac{R^2}{6}}}{\frac{\sinh(LR)}{R}} - 1
\end{aligned}$$

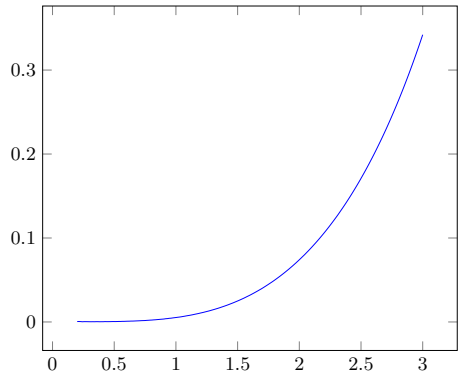


Fig. 7: Illustration of the maximum relative variations of $\mathbb{E}[Z_c]$, with the function $\Delta : x \mapsto \frac{e^{\frac{x^2}{6}}}{\frac{\sinh(x)}{x}} - 1$. The x-axis represents the radius considered and th e y- axis the value of the maximum relative variation.

This ratio does not depend on the dimension, regardless of the radius of the sphere considered. The graph of the function $\Delta : R \mapsto \Delta(R)$ for $L = 1$ is drawn in Figure 7. In particular, $||\Delta||_{\infty,[0,2]} \leq 10^{-1}$. In particular, this implies that if $R \leq 2$ (and $L \leq 1$):

$$\frac{\max_{c\in\mathcal{S}_R} \mathbb{E}[Z_c] - \min_{c\in\mathcal{S}_R} \mathbb{E}[Z_c]}{\min_{c\in\mathcal{S}_R} \mathbb{E}[Z_c]} = \Delta(R) \leq 10^{-1}$$

Finally, if Ω_η is replaced by the domain defined by

$$R \leq ||x||_2 \leq R + \eta$$

Then the extremum of f on Ω_η can be deduced from Lemma 2 and are given by

$$\min_{c\in\Omega_\eta} f(c) = L^{d-1}\frac{\sinh(LR)}{R}$$

$$\max_{c\in\Omega_\eta} f(c) = [\frac{\sqrt{d}}{R+\eta}\sinh(\frac{L(R+\eta)}{\sqrt{d}})]^d$$

Similarly,

$$\Delta(R) \underset{d\to+\infty}{\sim} L\frac{e^{\frac{(R+\eta)^2}{6}}}{\frac{\sinh(R)}{R}} - 1$$

Let $L = 1$ and denote by

$$\Delta_\eta := \frac{e^{\frac{(R+\eta)^2}{6}}}{\frac{\sinh(R)}{R}} - 1 \tag{43}$$

such maximum variation for a given η. Plots of Δ_η for several values of η are given in Fig. 1b.

Let Z_0 be a partition function for a constant discourse vector $x_0 \in \mathcal{S}_R$. The two events are equivalent:

$$\begin{aligned} &| Z_c - \mathbb{E}[Z_c] | > \epsilon \mathbb{E}[Z_c] \iff \\ &\quad \Big| \frac{Z_c}{\mathbb{E}[Z_0]} - \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \Big| > \epsilon \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \end{aligned} \tag{44}$$

Using the previous study, we know that

$$\Big| \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} - 1 \Big| \leq ||\Delta||_\infty$$

Which implies that

$$\epsilon \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \geq \epsilon(1 - ||\Delta||_\infty)$$

From Equation 44:

$$\begin{aligned} &| Z_c - \mathbb{E}[Z_c] | > \epsilon \mathbb{E}[Z_c] \\ &\implies \Big| \frac{Z_c}{\mathbb{E}[Z_0]} - \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \Big| > \epsilon(1 - ||\Delta||_\infty) \end{aligned}$$

Let $\mathcal{E}$ be the event corresponding to the right hand side. Then:

$$\begin{aligned} \mathbb{P}(\mathcal{E}) &\leq \mathbb{P}(|Z_c - \mathbb{E}[Z_c]| > \epsilon \mathbb{E}[Z_c]) \\ &\leq \alpha \end{aligned} \tag{45}$$

where the second line is obtained from Equation 36. We recall that ϵ is an arbitrarily small real number, and

$$\alpha = \exp\left(- \frac{\frac{1}{2}\varepsilon^2 n^2}{n\Lambda + \frac{1}{3}\sqrt{\Lambda}\varepsilon n} \right) \tag{46}$$

Hence, with (high) probability $1 - \alpha$:

$$\begin{aligned} -\epsilon(1 - ||\Delta||_\infty) + \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \;&\leq\; \frac{Z_c}{\mathbb{E}[Z_0]} \\ &\leq\; \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} + \epsilon(1 - ||\Delta||_\infty) \\ &\leq\; \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} + \epsilon(1 + ||\Delta||_\infty) \end{aligned} \tag{47}$$

Again, using:

$$1 - ||\Delta||_\infty \leq \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \leq 1 + ||\Delta||_\infty$$

We finally have with probability $1 - \alpha$:

$$(1 - \epsilon)(1 - ||\Delta||_\infty)\, \mathbb{E}[Z_0] \;\leq\; Z_c \;\leq\; (1 + \epsilon)(1 + ||\Delta||_\infty)\, \mathbb{E}[Z_0]$$

ϵ is arbitrarily small, and we saw that $||\Delta||_\infty \leq 10^{-1}$, for a domain close to a sphere of radius $R \leq 2$. Setting $\gamma = ||\Delta||_\infty$, this concludes the proof. □